Graphical Representation of Blood Group Data of Human Populations*

ROBERT BOUDREAU

Department of Mathematical Sciences
Virginia Commonwealth University
Richmond, VA 23284

and

C. RADHAKRISHNA RAO

Center for Multivariate Analysis
326 Classroom Building
Penn State University
University Park, PA 16802

ABSTRACT: A general survey of various measures of diversity within and distance between populations in gene frequencies of blood group systems is given. A method of ordering populations by an overall measure of diversity within and of clustering populations in terms of differences in pattern of diversity in different blood group systems is developed. Principal coordinate analysis and multidimensional scaling are used to represent populations with given distances between them graphically in an appropriate dimensional Euclidean space. Such graphical representations together with dendrograms are of great help in studying interrelationships between populations. The methods are illustrated using blood group data on some human populations. *AMS Classification index*: 62H30

KEY WORDS: cluster analysis, distance measures, diversity measures, multidimensional scaling, principal coordinate analysis, quadratic entropy, Rao's canonical coordinate analysis

1. INTRODUCTION

Blood group data is extensively used to study genetic differences between populations. The basic statistics used in such investigations are measures of "diversity within populations" and "distance between populations". For recent literature on the subject, reference may be made to papers by Carmelli and Cavalli-Sforza (1979), Karlin, Kenett and Bonné-Tamir (1979), Karlin, Carmelli and Bonné-Tamir (1982), Rao and Boudreau (1984) and Chakraborty and Rao (1991).

A diversity measure enables us to rank or group populations by homogeneity of individuals within a population, while a distance measure enables us to study interrelationships between populations and throw light on their evolution. Rao and Boudreau (1984) applied some recently developed methods in a study of differences in blood groups of jewish and gentile populations in Europe and the Middle East.

The purpose of the present communication is to develop graphical representations of the populations based on blood group data similar to the canonical coordinate plots in the continuous case introduced by Rao (1948, 1971).

We provide a general discussion of diversity and distance measures and choose particular measures based on Hellinger representation of multinomial distributions for illustrating the graphical representations.

For this purpose, we use the data on the antigenic blood group systems HLA-A, HLA-B, ABO, MNSs and Rh for 15 populations considered earlier by Karlin, Kenett and Bonné-Tamir (1979) and Rao and Boudreau (1984). A description of these fifteen populations in terms of Jewish-Gentile denominations and geographical locations is given in Table I.

TABLE I
Hierarchial classification of the populations.

	Denominations Jews-Gentiles	Historical-geographical H.G. groups	Individual populations
15 Populations	9[a] Jews (J)	4 Ashkenazi (A)	1 Polish (P′)
			1 Russian (R′)
			1 German (G′)
			1 Rumanian (Ru′)
		2 Shepardi (S)	1 Moroccan (M′)
			1 Libyan (L′)
		3 Oriental (O)	1 Iraqi (I′)
			1 Yemenite (Y′)
			1 Cochin (C′)
	6 Gentiles (G)	3 Middle Eastern (ME)	1 Arab (Ab)
			1 Armenian (Am)
			1 Samaritan (S)
		3 European (E)	1 German (G)
			1 Polish (P)
			1 Russian (R)
H^O	H^D	H^G	H^P

[a] Indicates the number of populations in a category at any given level of classification. H^O, H^D, H^G, H^P are within diversities at different levels.

2. ANALYSIS OF DIVERSITY WITHIN POPULATIONS

2.1. *Some known measures of diversity*

We consider the set of multinomial distributions

$$\mathbb{P} = \{\mathbf{p} = (p_1, \ldots, p_k): p_i \geq 0, \Sigma p_i = 1\}, \tag{2.1}$$

and define a function H on $\mathbb{P}$ as a measure of diversity if it satisfies the two conditions given in Rao (1982b):

(i) $H(\mathbf{p}) = 0$ if all the components of $\mathbf{p}$ are zero except one and > 0 otherwise.

(ii) $H(\cdot)$ is a concave functional on $\mathbb{P}$. As a consequence the diversity in a mixture of two populations is not smaller than the average of the diversities within individual populations.

The two conditions, however, do not specify a diversity function uniquely; other criteria such as easy interpretability in genetic terms may have to be used in determining an appropriate function. Examples of diversity functions that received wide applications in biology are entropy functions such as:

$$H_S(\mathbf{p}) = -\Sigma p_i \log p_i \qquad \text{(Shannon)}$$

$$H_\alpha(\mathbf{p}) = \frac{1 - \Sigma p_i^\alpha}{2^{\alpha-1} - 1} \qquad \text{(α-order entropy of Havrda and Charavát)}$$

$$H_P(\mathbf{p}) = -\Sigma[p_i \log p_i + (1 - p_i) \log(1 - p_i)] \qquad \text{(paired Shannon entropy)}$$

$$H_R(\mathbf{p}) = (1 - \alpha)^{-1} \log \Sigma p_i^\alpha \qquad \text{(α-degree entropy of Renyi)}$$

Recently, Rao (1982a, b, c, 1984) introduced a general diversity measure called the quadratic entropy

$$H_Q(\mathbf{p}) = \Sigma \Sigma d_{ij} p_i p_j, \tag{2.2}$$

where d_{ij} is a nonnegative number representing an intrinsic difference between the categories i and j: the H_Q then is the average difference between two individuals drawn at random from a population. In such a case, H_Q could be interpreted in terms of chosen numbers d_{ij}. Thus, if $p_1, \ldots, p_k$ are frequencies of different alleles of a gene at a locus on a chromosome and $d_{ij} = 1$ if $i \neq j$ and $d_{ii} = 0$, then

$$H_Q(\mathbf{p}) = 1 - \Sigma p_i^2 = H_2 \qquad \text{(Gini-Simpson index)}$$

which is the well-known index of gene diversity. If we consider p_i as the frequencies of genotypes in a random mating population and define d_{ij} as the proportion of genes not common to two genotypes i and j, then

$$H_Q(\mathbf{p}) = 1 - \Sigma p_i^2 - \Sigma p_i^2(1 - p_i)^2 = H_L,$$

which was introduced by Latter (1973). Other examples and necessary restrictions on d_{ij} to ensure the concavity of H_Q are given in Rao (1982b, 1982c).

2.2. *A new measure of diversity*

We start with the Hellinger representation of the multinomial distribution

$$(2.3) \qquad \mathbf{p} = (p_1, \ldots, p_k),$$

by a point

$$(2.4) \qquad (\sqrt{p_1}, \ldots, \sqrt{p_k}),$$

on the hypersphere in k dimensions. The maximum diversity is usually associated with the multinomial distribution where each p_i is equal to $1/k$, i.e., the point

$$(2.5) \qquad (1/\sqrt{k}, \ldots, 1/\sqrt{k}),$$

on the hypersphere. We may then define the diversity of (2.3) by a monotone decreasing function of the angle of separation between (2.4) and (2.5), such as the cosine function

$$(2.6) \qquad k^{-1/2}(\sqrt{p_1} + \ldots + \sqrt{p_k}).$$

For some technical reasons, we define the diversity by

$$(2.7) \qquad H_1(\mathbf{p}) = \sqrt{p_1} + \ldots + \sqrt{p_k},$$

by dropping the factor $k^{-1/2}$. The range of (2.7) is the interval $[1, \sqrt{k}]$, with unity representing complete homogeneity and $\sqrt{k}$ complete heterogeneity. We may transform H_1 in (2.7) to

$$(2.8) \qquad H_2(\mathbf{p}) = \frac{H_1(\mathbf{p}) - 1}{\sqrt{k} - 1},$$

which has the range [0, 1] with zero representing complete homogeneity. We may also transform H_1 in (2.7) to

$$(2.9) \qquad H_3(\mathbf{p}) = \log H_1(\mathbf{p})$$

which has the range $[0, \log \sqrt{k}]$. It is seen that $H_2(\mathbf{p})$ and $H_3(\mathbf{p})$ belong respectively to the classes $H_\alpha(\mathbf{p})$ and $H_\beta(\mathbf{p})$ defined above.

The expressions (2.7), (2.8) and (2.9) are possible measures of diversity when we are considering a single blood group system. If there are m systems with associated vectors $\mathbf{p}_1, \ldots, \mathbf{p}_m$ of sizes $k_1, \ldots, k_m$, then a composite measure of diversity may be defined as

$$(2.10) \qquad H_i\,(\mathbf{p}_1, \ldots, \mathbf{p}_m) = H_i(\mathbf{p}_1) + \ldots + H_i(\mathbf{p}_m)$$

for any chosen H_i, $i = 1, 2, 3$. In our study, we chose H_2 for the computation of overall diversity

NOTE 1. The diversity measure $H_2(\mathbf{p})$ has also the representation

$$(2.11) \qquad H_2(\mathbf{p}) = 1 - \frac{\|(\sqrt{p_1}, \ldots, \sqrt{p_k}) - (1/\sqrt{k}, \ldots, 1/\sqrt{k})\|^2}{\max\limits_{\mathbf{p}} \|(\sqrt{p_1}, \ldots, \sqrt{p_k}) - (1/\sqrt{k}, \ldots, 1/\sqrt{k})\|^2}$$

where $\|\bullet\|$ is the Euclidean norm. Thus, diversity is measured by how far (2.4) is from (2.5), with small distances indicating higher diversity.

NOTE 2. It is seen that the measures H_1, H_2, and H_3 have the required properties:

(i) $H(\mathbf{p}) = 0$ iff one $p_i = 1$ and the rest are zero.
(ii) $H(\mathbf{p}) > 0$ otherwise and attains the maximum value when all p_i are equal
(iii) $H(\mathbf{p})$ is a concave functional over the space of multinomial distributions with k alleles.

(See Burbea and Rao (1982a, 1982b) and Lewontine (1972) for a discussion of concave diversity functionals.)

NOTE 3. Let $\mathbf{p} = (p_1, \ldots, p_k)$, $\mathbf{q} = (q_1, \ldots, q_r)$ be vectors of frequencies corresponding to two blood group systems which are independently inherited. Represent the joint distribution under independence by

$$(\mathbf{p}, \mathbf{q}) = \{p_i q_j;\ i = 1, \ldots, k;\ j = 1, \ldots, r\}.$$

Then:

$$H_1(\mathbf{p}, \mathbf{q}) \geq \lambda H_1(\mathbf{p}) + (1 - \lambda)H_1(\mathbf{q}),\ 0 \leq \lambda \leq 1.$$

$$H_2(\mathbf{p}, \mathbf{q}) = H_2(\mathbf{p}) + H_2(\mathbf{q}).$$

2.3. *Ordering of populations by diversity*

Table II gives the values of diversity within a population as defined in (2.8) for each population separately for each blood group system and overall which is the average over all blood group systems. Judging from the overall diversity values, the general conclusion is that the Samaritans (S) have the lowest, the Yemenite (Y′) and Cochin (C′) jews have the next lowest and the rest have nearly equal but a higher degree of diversity. The same order is maintained generally for individual blood group systems. A similar conclusion was drawn using other measures of diversity in Rao and Boudreau (1984).

The concavity of a diversity functional enables us to decompose the total diversity (T) in all the populations put together as between (B) and within (W) populations and compute the percentage of diversity due to differences between populations. The decomposition is obtained as follows.

TABLE II
Within and between population diversities by blood group systems.

Population	HLA-A	HLA-B	ABO	MNS	h	Overall
			Diversity (H_2)			
P′	0.9579	0.9047	0.6871	0.9525	0.6168	0.8238
R′	0.9412	0.9358	0.7404	0.9537	0.6547	0.8451
G′	0.9561	0.8834	0.6706	0.9365	0.6385	0.8170
Ru′	0.9539	0.9199	0.6833	0.9658	0.6366	0.8319
M′	0.9064	0.9461	0.6492	0.9419	0.6017	0.8091
L′	0.9396	0.9473	0.6556	0.9556	0.5701	0.8136
I′	0.9387	0.8820	0.7094	0.9812	0.6986	0.8420
Y′	0.8567	0.8691	0.6480	0.8822	0.5789	0.7670
C′	0.9653	0.8474	0.5164	0.9181	0.4924	0.7479
Ab	0.9336	0.8772	0.6516	0.9723	0.7608	0.8391
Am	0.9509	0.8840	0.7076	0.9550	0.5744	0.8144
S	0.7337	0.5368	0.5957	0.9861	0.5003	0.6705
G	0.9280	0.9201	0.6343	0.9219	0.6014	0.8012
P	0.8955	0.9296	0.6867	0.9402	0.6758	0.8255
R	0.8838	0.9627	0.6844	0.9231	0.6593	0.8227
			All populations			
Between	0.0383 (4.0%)	0.0749 (7.8%)	0.0278 (4.0%)	0.0110 (1.2%)	0.0614 (9.0%)	0.0427 (5.0%)
within	0.9161	0.8831	0.6613	0.9457	0.6174	0.8047
Total	0.9544	0.9580	0.6891	0.9568	0.6788	0.8474
			Omitting S			
Between	0.0280 (2.9%)	0.0578 (6.0%)	0.0244 (3.5%)	0.0089 (0.9%)	0.0594 (8.7%)	0.0357 (%)
Within	0.9292	0.9078	0.6660	0.9429	0.6257	0.8143
Total	0.9571	0.9656	0.6905	0.9518	0.6851	0.8500

Let us denote the gene frequencies in any blood group system for the i-th population by

$$(2.12)\quad \mathbf{p}_i = (p_{1i}, \ldots, p_{ki}),\ i = 1, \ldots, 15,$$

and the average over all populations by

$$(2.13)\quad \mathbf{p} = (p_1, \ldots, p_k),\quad p_i = \sum_{i}^{15} w_i p_{ji},$$

where w_i is the weight given to the i-th population, which is usually the proportion of individuals in the i-th population to the total over all populations. (In our study we have given equal weight to all the populations). If we denote diversity as a functional H(•) over the space of multinomial distributions, then the desired decomposition can be obtained as

$$(2.14)\quad \mathrm{T} = \mathrm{B} + \mathrm{W},$$

where

$$\mathrm{T} = \mathrm{H}(\mathbf{p}),\ \mathrm{W} = \sum w_i\ \mathrm{H}(\mathbf{p}_i).$$

The value of B is obtained as the difference T – W, which is called the Jensen difference (see Rao (1982a)).

Table II gives also the values of B and W for each blood group system and over all the systems, considering all the populations and also after omitting S which has the least within diversity. It is seen that the percentage of diversity between populations to total varies from about 1 to 9 percent over different blood group systems. The highest values are associated with Rh and HLA–B systems, which provide the maximum discrimination between populations.

2.4. *Clustering of populations by diversity*

In section 2.3, a linear ordering of the populations was obtained in terms of the pooled diversity over all the blood group systems. We can, however, differentiate between populations by constructing distances between populations based on the pattern of diversities in individual blood group systems. For instance, if $d_1, \ldots, d_5$ and $d'_1, \ldots, d'_5$ are the diversities in five blood group systems as recorded in Table II for two populations, say i and j, then the dissimilarity between the two populations i and j in terms of the pattern of diversities in different blood group systems may be defined as

$$(2.15)\quad d_{ij} = \sum_{r=1}^{5} \left(\frac{d_r + d'_r}{2\sqrt{d_r d'_r}} - 1 \right),$$

which lies in range $[0, \infty)$. A different measure of dissimilarity was used in Rao and Boudreau (1984).

TABLE III

Dissimilarities between populations in the pattern of within population diversities in different blood group systems.

	P′	R′	G′	Ru′	M′	L′	I′	Ab	Am	G	P	R	Y′	C′
R′	13.2													
G′	3.3	17.9												
Ru′	1.9	9.8	3.7											
M′	11.3	32.6	15.2	12.3										
L′	13.6	42.6	23.6	18.8	5.6									
I′	23.1	13.0	17.2	15.4	47.5	66.7								
Ab	61.2	54.4	42.0	46.0	78.4	112.1	18.3							
Am	8.2	28.2	18.1	16.9	20.9	13.5	49.1	108.3						
G	11.7	41.0	11.8	14.6	2.9	7.8	51.0	76.5	21.9					
P	17.3	11.8	13.4	10.5	21.4	42.5	11.2	28.8	42.2	27.1				
R	19.7	15.1	19.0	13.9	15.6	35.2	24.6	46.6	42.4	23.3	2.9			
Y′	34.3	66.6	33.3	43.5	20.2	28.4	79.3	114.7	31.6	16.9	47.4	41.7		
C′	172.6	279.1	172.6	192.5	136.7	116.5	288.0	312.1	158.3	113.3	245.7	236.0	117.7	
S	513.4	617.3	494.8	547.9	514.6	516.6	563.7	606.4	459.0	487.3	570.3	597.7	372.6	388.1

The values in the table are obtained by using the formula (2.15) and multiplying the result by 10^4.

The values of (2.15) for all pairs of populations are given in Table III. It is seen that in terms of diversity, the populations S and C′ are distinct from the rest. The dendrogram for all the populations based on the matrix of dissimilarities (Table III), using the method of complete linkage is given in Figure 1.

From the dendrogram in Figure 1, it is seen that there are four different clusters

S′ C′, (Y′, L′, M′, G, Am, G′, P′, Ru′), (Ab, P, R, I′, R′)

with different patterns of within population diversities. Within the third cluster Y′ seems to have a somewhat different pattern of diversity than the rest and within the fourth cluster Ab seems to differ slightly from the rest.

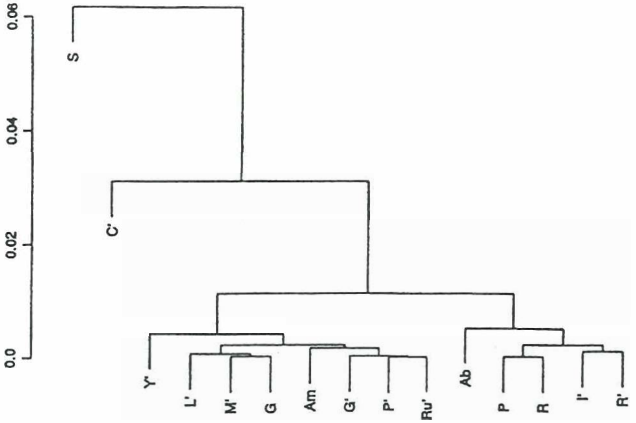

Fig. 1. Dendrogram (complete linkage) based on the dissimilarity matrix (Table I) in diversities.

3. INTERRELATIONSHIPS BETWEEN POPULATIONS

3.1. *Distance measures*

All studies of interrelationships between populations start with a matrix of similarities or dissimilarities (distances) between populations. In the present context, there are various ways of computing dissimilarities or distances.

Let the gene frequencies at the i-th locus in population π_α be

$$\text{(3.1)} \quad \mathbf{p}_i^\alpha = (p_{i1}^\alpha, \ldots, p_{ik_i}^\alpha), \; i = 1, \ldots, m; \; \alpha = 1, \ldots, N,$$

where m is the number of blood group systems and N is the number of populations. The whole gametic array can be represented by the partitioned vector

$$(3.2) \quad \mathbf{p}^\alpha = (\mathbf{p}_1^\alpha, \ldots, \mathbf{p}_m^\alpha), \ \alpha = 1, \ldots, N.$$

Some examples of distance functions based on (3.2) are given below.

(i) Nei's minimum distance

$$D_{\alpha\beta}^{(1)} = (\delta_{\alpha\beta}^1)^2 = (\mathbf{p}^\alpha - \mathbf{p}^\beta)(\mathbf{p}^\alpha - \mathbf{p}^\beta)' = \sum_{i=1}^{m} \sum_{n=1}^{k_i} (p_{ir}^\alpha - p_{ir}^\beta)^2. \tag{3.3}$$

(ii) Nei's standard distance

$$D_{\alpha\beta}^{(2)} = -\log \cos \theta_{\alpha\beta}, \tag{3.4}$$

where

$$\cos \theta_{\alpha\beta} = \sum_1^m \mathbf{p}_i^\alpha(\mathbf{p}_i^\beta)' \Big/ \left[\sum_1^m \mathbf{p}_i^\alpha(\mathbf{p}_i^\alpha)'\right]^{1/2} \left[\sum_1^m \mathbf{p}_i^\beta(\mathbf{p}_i^\beta)'\right]^{1/2}.$$

$D_{\alpha\beta}^{(2)}$ as defined in (3.4) does not satisfy the postulates of a distance function. However, we may use the angle

$$\theta_{\alpha\beta}^{(2)} = \cos^{-1}[\exp(-D_{\alpha\beta}^{(2)})], \tag{3.5a}$$

or the chord length

$$\delta_{\alpha\beta}^{(2)} = 2 \sin(\theta_{\alpha\beta}^{(2)}/2), \tag{3.5b}$$

which are distance functions.

(iii) Nei's maximum distance

$$D_{\alpha\beta}^{(3)} = -\log\left(\prod_{i=1}^{m} \cos \theta_{\alpha\beta}^{(i)}\right) \tag{3.6}$$

where

$$\cos \theta_{\alpha\beta}^{(i)} = \mathbf{p}_i^\alpha(\mathbf{p}_i^\beta)'/[\mathbf{p}_i^\alpha(\mathbf{p}_i^\alpha)']^{1/2}[\mathbf{p}_i^\beta(\mathbf{p}_i^\beta)']^{1/2}.$$

Instead of $D_{\alpha\beta}^{(3)}$, we may use the angle

$$\theta_{\alpha\beta}^{(3)} = -\cos^{-1}\left(\prod_{i=1}^{m} \cos \theta_{\alpha\beta}^{(i)}\right), \tag{3.7a}$$

or the chord length

$$\delta_{\alpha\beta}^{(3)} = 2 \sin(\theta_{\alpha\beta}^{(2)}/2), \tag{3.7b}$$

which are distance functions. (See Nei (1973), (1978) for a description and use of the distance or dissimilarity functions introduced by him.)

If we denote

$$\mathbf{p}_i^\alpha = (\sqrt{p_{i1}^\alpha}, \ldots, \sqrt{p_{ik_i}^\alpha}), \tag{3.8}$$

and apply the formula (3.3), we get what is known as Matusita (1957) or Hellinger distance

$$\text{(iv)} \qquad (\delta_{\alpha\beta}^{(4)}) = \left[\sum_{i=1}^{m} \sum_{r=1}^{k_i} (\sqrt{p_{ir}^\alpha} - \sqrt{p_{ir}^\beta})^2\right]^{1/2}. \tag{3.9}$$

Applying the formula (3.6), we get

$$\text{(v)} \qquad D_{\alpha\beta}^{(5)} = -\log\left(\prod_{i=1}^{m} \cos \theta_{\alpha\beta}^{(i)}\right), \tag{3.10}$$

where

$$\cos \theta_{\alpha\beta}^{(i)} = \sum_{r=1}^{k_i} (p_{ir}^\alpha p_{ir}^\beta)^{1/2}.$$

Instead of $D_{\alpha\beta}^{(5)}$, we may use the angle

$$\theta_{\alpha\beta}^{(5)} = \cos^{-1} \prod_{i=1}^{m} \cos \theta_{\alpha\beta}^{(i)} \tag{3.11a}$$

or the chord length

$$(\delta_{\alpha\beta}^{(5)}) = 2 \sin(\theta_{\alpha\beta}^{(5)}/2), \tag{3.11b}$$

which are distance functions. We call $\delta_{\alpha\beta}^{(5)}$ the composite chord distance.

In our previous study (Rao and Boudreau (1984)), we used Nei's distance functions. In the present study we use Matusita distance $\delta_{\alpha\beta}^{(4)}$ and the chord distance $\delta_{\alpha\beta}^{(5)}$ based on the Hellinger representation of gene frequencies.

Tables IV and V give respectively the values of Matusita ($\delta_{\alpha\beta}^{(4)}$) and composite chord ($\delta_{\alpha\beta}^{(5)}$) distances between all pairs of populations. The corresponding dendograms using the complete linkage method are given in Figures 2 and 3. These dendograms seem to indicate close clustering of the European gentiles and so also the European (Akshanazi) jews

(G, P, R) and (G′, Ru′, P′, R′)

with some separation between the two clusters. The Iraqi (I′), Moroccan (M′), Yemenite (Y′), Cochin (C′) and Lybian (L′) jews seem to be a little distant from the other jews and also from the European gentiles. The Samaritans (S), Arabs (Ab) and Armenians (Am) take distinct positions separated among themselves and from the other groups of populations.

In order to study the interrelationships further and to provide a graphical representation of the populations based on the distances between them, we have used the method of principal coordinate analysis (PCA) and non-

TABLE IV
The matrix of Matusita distances (3.9) between populations.

	P′	R′	G′	Ru′	M′	L′	I′	Ab	Am	G	P	R	Y′	C′
P′														
R′	0.243													
G′	0.316	0.373												
Ru′	0.249	0.286	0.380											
M′	0.519	0.510	0.538	0.508										
L′	0.527	0.465	0.569	0.509	0.571									
I′	0.573	0.570	0.626	0.591	0.516	0.544								
Ab	0.695	0.649	0.747	0.723	0.700	0.637	0.703							
Am	0.631	0.661	0.633	0.658	0.705	0.585	0.628	0.667						
G	0.595	0.604	0.582	0.588	0.569	0.526	0.634	0.756	0.581					
P	0.602	0.598	0.552	0.606	0.560	0.539	0.601	0.734	0.561	0.356				
R	0.571	0.520	0.588	0.561	0.538	0.474	0.577	0.692	0.607	0.445	0.335			
Y′	0.625	0.679	0.692	0.642	0.685	0.699	0.756	0.810	0.747	0.729	0.705	0.745		
C′	0.644	0.696	0.660	0.677	0.692	0.700	0.763	0.869	0.762	0.586	0.628	0.716	0.637	
S	0.934	0.947	0.936	0.956	1.010	0.943	1.029	0.983	0.929	0.994	0.978	1.030	0.975	1.042

TABLE V
The matrix of composite chord distances (3.11b) between populations.

	P′	R′	G′	Ru′	M′	L′	I′	Ab	Am	G	P	R	Y′	C′
P′														
R′	0.242													
G′	0.313	0.369												
Ru′	0.248	0.283	0.377											
M′	0.507	0.499	0.527	0.498										
L′	0.516	0.457	0.557	0.498	0.554									
I′	0.559	0.555	0.610	0.576	0.503	0.530								
Ab	0.669	0.630	0.716	0.695	0.672	0.618	0.674							
Am	0.613	0.640	0.618	0.640	0.674	0.570	0.611	0.647						
G	0.581	0.587	0.571	0.574	0.552	0.515	0.612	0.725	0.567					
P	0.587	0.581	0.544	0.592	0.545	0.526	0.583	0.704	0.547	0.352				
R	0.558	0.509	0.577	0.548	0.525	0.465	0.561	0.666	0.589	0.439	0.332			
Y′	0.603	0.651	0.664	0.620	0.656	0.667	0.716	0.765	0.708	0.698	0.675	0.711		
C′	0.623	0.667	0.639	0.655	0.665	0.672	0.725	0.817	0.728	0.568	0.606	0.687	0.618	
S	0.874	0.886	0.875	0.897	0.933	0.879	0.939	0.902	0.866	0.928	0.910	0.948	0.909	0.965

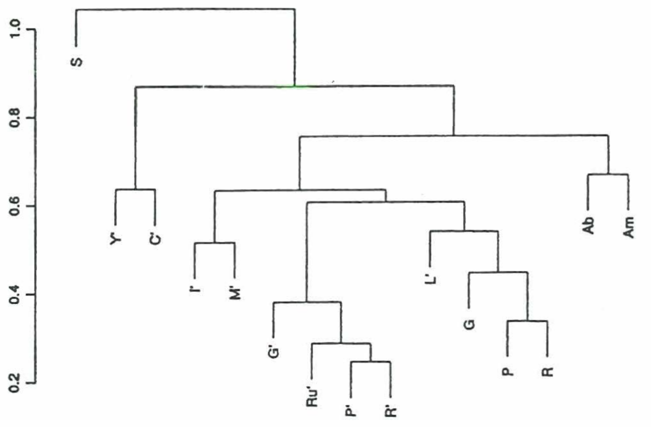

Fig. 2. Dendrogram (complete linkage) based on Matusita distances.

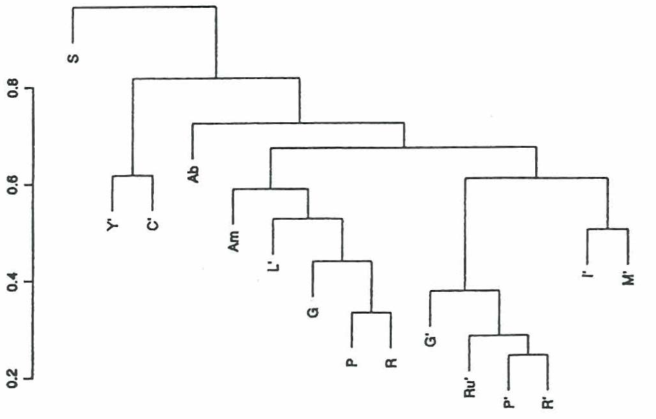

Fig. 3. Dendrogram (complete linkage) based on the composite chord distances.

metric multidimensional scaling (NMMDS). A brief description of these methods is given in the Appendix. For purposes of illustration, we chose the Matusita distance matrix as the basis for graphical representation. The Samaritans (S) are omitted from the analysis as they are quite distant from the other populations and their inclusion might distort the relative positions of the other populations in the lower dimensional graphical representations. As a practical rule, we suggest the following. First, construct a dendrogram and/or use PCA (or NMMDS) on the matrix of distances considering all the populations. Then pick up broad and widely separated clusters by examining the dendograms and the PCA (or NMMDS) plots. Then apply PCA or NMMDS separately on each cluster to obtain a graphical representation of populations within each cluster.

Principal coordinate analysis: The 15 × 15 matrix D of distances is transformed into the 15 × 15 matrix B as explained in the Appendix and its spectral decomposition is obtained in the form

$$\text{(3.13)} \quad \mathbf{B} = \lambda_1^2\mathbf{P}_1\mathbf{P}_1' + \lambda_2^2\mathbf{P}_2\mathbf{P}_2' + \ldots$$

where $\lambda_1^2, \lambda_2^2, \ldots$ are the eigen values and $\mathbf{P}_1, \mathbf{P}_2, \ldots$ are the corresponding eigen vectors of B. We have to choose the appropriate dimension for representing the 15 populations. This is done by examining the ratios

$$\text{(3.14)} \quad \frac{\lambda_1^2}{\Sigma\lambda_i^2}, \quad \frac{\lambda_1^2 + \lambda_2^2}{\Sigma\lambda_i^2}, \ldots,$$

which in the present case turn out to be in terms of percentages

$$\text{(3.15)} \quad 20.2\%, 39.4\%, 54.6\%, 64.8\%, 73.3\%, 80.8\%, \ldots.$$

The number of dimensions needed for graphical representation to capture most of the differences between populations is judged by the ratios in (3.14). If the second ratio in (3.14) is large, then a two dimensional representation is adequate. Otherwise, we may have to consider a higher dimensional representation. In the present example, it may be necessary to go up to five dimensions which explain about 73% of the differences between populations. The coordinates associated with the first five dimensions are given in Table VI. Figure 4 gives the plot of actual (Matusita) distances versus the distances in the five dimensional reduced space of principal coordinates. The association seems to be fairly satisfactory. Using the principal coordinates, biplots are made for every pair of coordinates as shown in Figures 6.1–6.10. The broad conclusions form these plots, which we call the grand tour, are as follows.

The population sets (G, P, R) and (G′, P′, R′, Ru′), are separated, but within each set the populations stick together in all the plots confirming two closely knit clusters. C′, Y′, Am, Ab wander around keeping some distances among themselves and not associate with any other particular population. I′, M′, L′ behave in the same way though not so widely separated as C′, Y′, Am and Ab.

The distinctions between I′, M′ and L′ populations are brought out clearly in the 4-th and 5th dimensions. It is interesting to note that the first dimension clearly separates the jews (except L′ and I′) and the gentiles while

Table VI
The principal coordinates in 5 dimensions obtained from the distance matrix in Table IV with S omitted.

	X1	X2	X3	X4	X5
P	0.1947	–0.1839	0.0509	–0.0846	0.0149
R′	0.1160	–0.2333	0.0756	–0.0490	–0.0632
G′	0.1669	–0.1195	0.1282	–0.1393	0.0147
Ru′	0.1860	–0.1833	0.0876	–0.0561	0.0050
M′	0.0208	–0.0581	0.0693	0.2743	–0.0298
L′	–0.1125	–0.0477	0.0148	0.0016	–0.0167
I′	–0.1325	–0.0946	0.0259	0.2620	0.2295
Ab	–0.2904	–0.2106	–0.3527	–0.0156	–0.2245
Am	–0.2356	0.0655	–0.1183	–0.2474	0.2560
G	–0.1049	0.2493	0.1445	–0.0443	–0.0638
P	–0.1482	0.2308	0.1391	–0.0170	–0.0289
R	–0.1864	0.0811	0.1651	0.0365	–0.0801
Y′	0.2764	0.1504	–0.3419	0.0792	0.1005
C′	0.2498	0.3539	–0.0879	–0.0003	–0.1137
%	20.2*	39.4	54.6	64.8	73.3

* The figures represent the percentage of differences explained by principal coordinates in different dimensions.

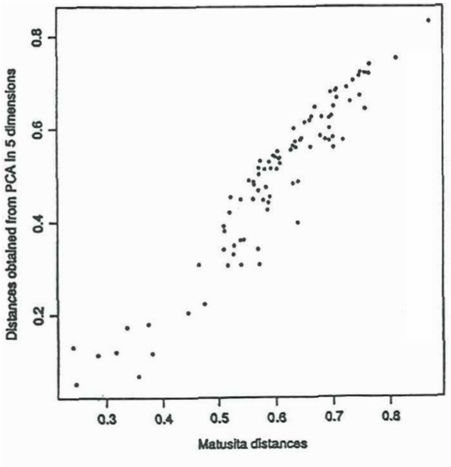

Fig. 4. Plot of Matusita distances from Table IV vs fitted distances using Principal Coordinates in 5 dimensions from Table VI.

Table VII
The Non-Metric Multidimensional Scaling coordinates in 5 dimensions obtained from the distance matrix in Table IV with S omitted.

	X1	X2	X3	X4	X5
P′	–0.0805	0.3503	0.2633	–0.2514	–0.1421
R′	0.1698	0.2180	0.4177	–0.0180	–0.3326
G′	–0.2242	–0.0090	0.4165	–0.4918	–0.3506
Ru′	–0.1288	0.2826	0.5000	0.0231	–0.1932
M′	–0.0625	0.0795	0.5550	0.4248	0.3542
L′	0.4479	–0.0960	–0.0221	0.2022	–0.3610
I′	0.5404	–0.1533	0.5059	–0.2165	0.6502
Ab	1.2933	0.7497	–0.6495	0.3245	–0.0316
Am	0.4594	–0.4612	–0.7784	–0.6035	0.0426
G	–0.2724	–0.5974	–0.2575	0.2809	–0.1092
P	–0.1611	–0.6936	–0.0976	–0.0425	0.2515
R	0.3122	–0.5731	0.0060	0.2645	–0.0290
Y′	–0.8732	1.0716	–0.4228	–0.1458	0.3135
C′	–1.4204	–0.1682	–0.4364	0.2494	–0.0626

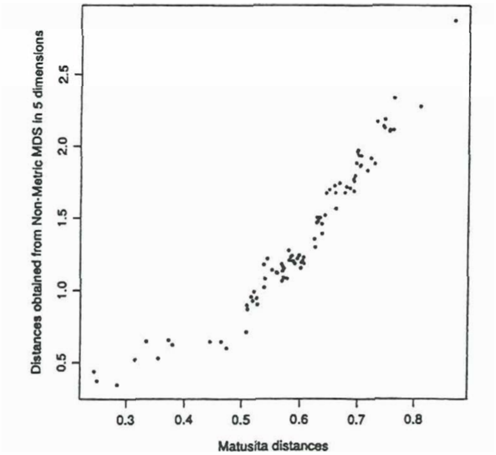

Fig. 5. Plot of Matusita distances from Table IV vs fitted distances using Non-Metric Multidimensional Scaling in 5 dimensions from Table VII.

the second dimension separates the Cochin (C′) and Yemenite (Y′) jews from the rest of the jews, and also the European jews from the European gentiles.

Non-metric multidimensional scaling: Figures 7.1–7.10 provide the grand tour in five dimensions using the non-metric multidimensional scaling program. The coordinates in the five dimensional space are given in Table VII. The general conclusions about the interrelationships between populations are the same as in the case of PCA. Figure 5 gives the plot of actual (Matusita) distances versus the distances in the five dimensional space determined by non-metric multidimensional scaling. Again the association seems to be fairly satisfactory.

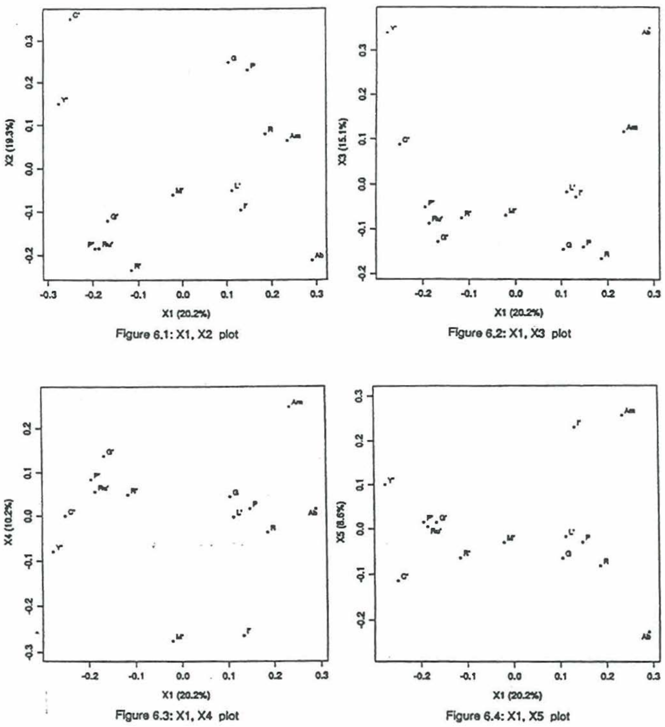

Figure 6.1: X1, X2 plot

Figure 6.2: X1, X3 plot

Figure 6.3: X1, X4 plot

Figure 6.4: X1, X5 plot

Fig. 6. Grand tour of the two dimensional plots using the 5 dimensional principal coordinates in Table VI.

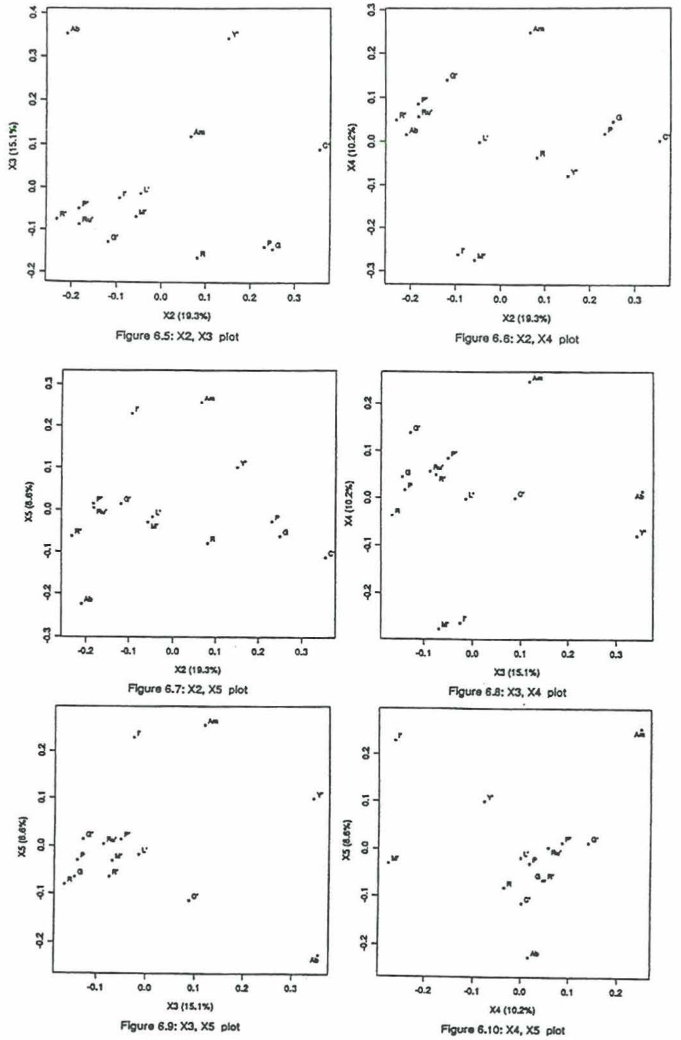

Figure 6.5: X2, X3 plot

Figure 6.6: X2, X4 plot

Figure 6.7: X2, X5 plot

Figure 6.8: X3, X4 plot

Figure 6.9: X3, X5 plot

Figure 6.10: X4, X5 plot

Fig. 6. *(Continued).*

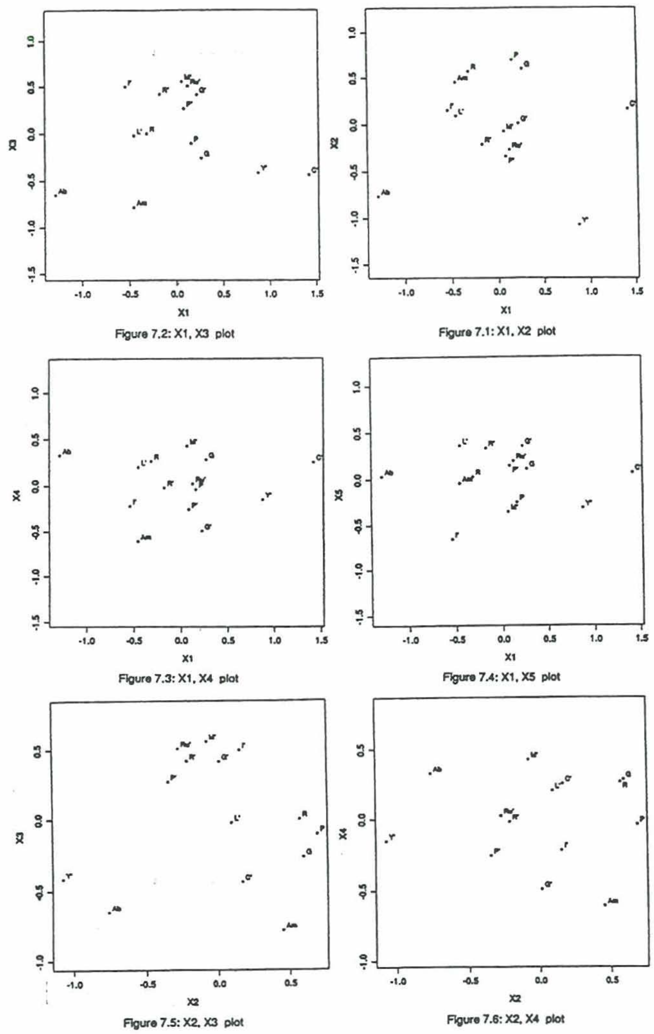

Figure 7.2: X1, X3 plot

Figure 7.1: X1, X2 plot

Figure 7.3: X1, X4 plot

Figure 7.4: X1, X5 plot

Figure 7.5: X2, X3 plot

Figure 7.6: X2, X4 plot

Fig. 7. Grand tour of the two dimensional plots using the 5 Non-Metric Multidimensional Scaling coordinates in Table VII.

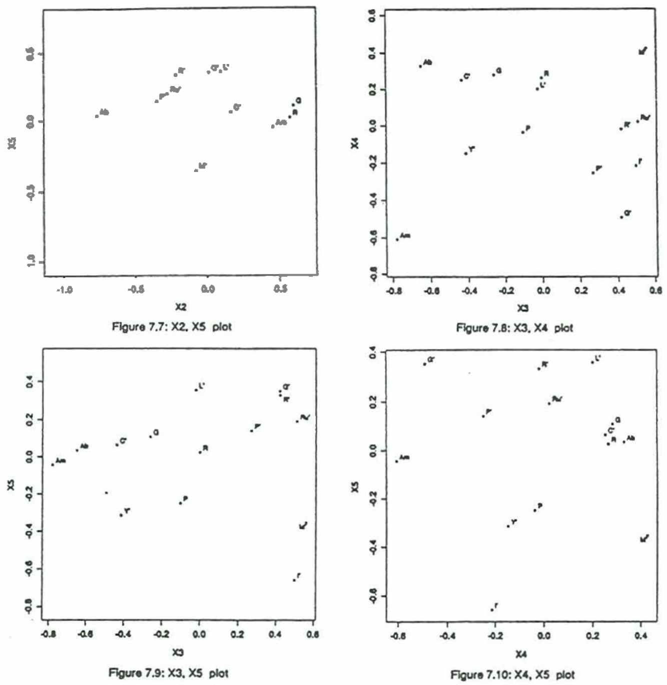

Figure 7.7: X2, X5 plot

Figure 7.8: X3, X4 plot

Figure 7.9: X3, X5 plot

Figure 7.10: X4, X5 plot

Fig. 7. *(Continued).*

APPENDIX

In taxonomic investigations we have a given matrix of dissimilarities between populations (or taxonomic units) and the problem is to represent the populations in an appropriate low dimensional Euclidean space such that the resulting configuration of points is consistent with (or reflect) the original dissimilarities. We describe two methods of doing this depending on the nature of the dissimilarity measures used.

Principal coordinate analysis: A description of this method, proposed by Torgerson (1952), is given in Rao (1964). It is applicable ideally in situations where the dissimilarities between populations can be expressed as distances between points representing the populations in a Euclidean space of a certain number of dimensions. If this dimension is large, we may wish to find a representation of the populations in a lower dimensional Euclidean space such that the difference between the configurations of points in the original and reduced spaces is as small as possible. Let d_{ij}^2 be the squared distance between populations i and j in the original space and represent by Δ the matrix

$$\text{(A.1)} \quad \Delta = (d_{ij}^2).$$

From Δ, we derive the matrix

$$\text{(A.2)} \quad \mathrm{B} = -\frac{1}{2}(\mathrm{I} - n^{-1}\mathbf{1}\,\mathbf{1}')\Delta(\mathrm{I} - n^{-1}\mathbf{1}\,\mathbf{1}'),$$

where n is the number of populations, I is the $n \times n$ unit matrix and $\mathbf{1}$ is the n-vector of unities, and obtain the spectral decomposition of B

$$\text{(A.3)} \quad \mathrm{B} = \lambda_1^2\mathbf{P}_1\mathbf{P}_1' + \ldots + \lambda_n^2\mathbf{P}_n\mathbf{P}_n'$$

Note that B is an $n \times n$ matrix, $\lambda_1^2, \ldots, \lambda_n^2$ are the eigen values of B and $\mathbf{P}_1, \ldots, \mathbf{P}_n$ are the corresponding eigen vectors. Then the coordinates for representing the populations in the best k-dimensional space are given by the rows of the matrix

$$\text{(A.4)} \quad \mathrm{B}_{(k)} = (\lambda_1\mathbf{P}_1, \lambda_2\mathbf{P}_2, \ldots, \lambda_k\mathbf{P}_k).$$

We have to make a choice of k depending on the magnitude of the ratios

$$\text{(A.5)} \quad \frac{\lambda_1^2 + \ldots + \lambda_k^2}{\lambda_1^2 + \ldots + \lambda_n^2}, \quad k = 1, 2, \ldots$$

The larger the ratio, the better is the representation of the populations in the lower dimensional space.

Though not strictly appropriate, the PCA can be used on any dissimilarity matrix, but the success of the method depends on the extent to which the dissimilarity matrix can be approximated by a matrix of squared Euclidean distances.

Multidimensional Scaling: A description of this method can be found in the book on *Multidimensional Scaling* by Kruskal and Wish (1978) who are the principal contributers to this area. The method can be applied to any kind of dissimilarity matrix. Let $\mathrm{D} = (d_{ij})$ be a given dissimilarity matrix and $\mathrm{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_n)$ be a $k \times n$ matrix with the i-th column vector providing the coordinates of the i-th population in a k-dimensional Euclidean space. The distance between populations i and j in this space is

$$\text{(A.6)} \quad d_{ij}^* = [(\mathbf{X}_i - \mathbf{X}_j)'(\mathbf{X}_i - \mathbf{X}_j)]^{1/2}.$$

To determine the consistency between the configurations determined by (d_{ij}) and (d_{ij}^*), we define what is called a stress function

$$\text{(A.7)} \quad \mathrm{S}(\mathrm{X}, f) = \frac{\sum\sum(d_{ij}^* - f(d_{ij}))^2}{\sum\sum(d_{ij}^*)^2}.$$

where f is a monotonic function. The problem of multidimensional scaling is that of minimizing $\mathrm{S}(\mathrm{X}, f)$ with respect to f and X. Suppose that the minimum is attained at X^*, f^*. Then X^*, the X associated with the minimum value of $\mathrm{S}(\mathrm{X}, f)$, gives the coordinates for the best possible representation of the populations in a k-dimensional Euclidean space.

The adequacy of fit is judged by the resulting stress $\mathrm{S}(\mathrm{X}^*, f^*)$. This value decreases with increase in k, and in practical work a judgement has to be made on the choice of k based on the stress value. Some guidelines for this purpose can be found in Kruskal and Wish (1978).

NOTE

* Research sponsored by the Air Force Office of Scientific Research under Grant AFOSR-91-0242.

REFERENCES

Burbea, J. and C. Radhakrishna Rao 1982a On the Convexity of Some Divergence Measures Based on Entropy Functions. IEEE Trans. Inform. Theor. 28: 489–495.

Burbea, J. and C. Radhakrishna Rao 1982b On the Convexity of Higher Order Jensen Differences Based on Entropy Functions. IEEE Trans. Inform. Theor. 28: 961–963.

Carmelli, D. and I. I. Cavalli-Sforza 1979 The Genetic Origin of the Jews: A Multivariate Approach. Hum. Biol. 51: 41–61.

Chakraborty, R. and C. Radhakrishna Rao 1991 Measurement of Genetic Variation for Evolutionary Studies. Handbook of Statistics, Vol. 8, North-Holland, pp. 271–316.

Karlin, S., R. Kenett and B. Bonné-Tamir 1979 Analysis of Biochemical and Genetic Data on Jewish Populations: II. Results and Interpretations of Heterogeneity Indices and Distance Measures with Respect to Standards. Am. J. Hum. Genet. 31: 341–365.

Karlin, S., D. Carmelli and B. Bonné-Tamir 1982 Analysis of Biochemical and Genetic Data on Jewish Populations: III. The Application of Individual Phenotypic Measurements for Population Comparisons. Am. J. Hum. Genet. 34: 50–64.

Kruskal, J. B. and M. Wish 1978 Multidimensional Scaling. Sage Publications, Berverly Hills, CA.

Latter, B. D. H. 1973 Measures of Genetic Distance Between Individuals and Populations. In Genetic Structure of Populations. N. E. Morton, ed., pp. 27–39. Univ. of Hawaii Press.

Lewontin, R. C. 1972 The Apportionment of Human Diversity. Evol. Biology 6: 381–398.

Matusita, K. 1957 Decision Rule Based on the Distance for the Classification Problem. Ann. Inst. Stat. Math. 8: 67–77.

Nei, M. 1973 Analysis of Gene Diversity in Subdivided Populations. Proc. Nat. Acad. Sci. 70: 3321–3323.

Nei, M. 1978 The Theory of Genetic Distance and Evolution of Human Races. Japan J. Human Genet. 23: 341–369.

Rao, C. Radhakrishna 1948 The Utilization of Multiple Measurements in Problems of Biological Classification. J. Roy. Statist. Soc. 10: 159–203.

Rao, C. Radhakrishna 1964 The Use and Interpretation of Principal Component Analysis in Applied Research. Sankhyā A 26: 329–358.

Rao, C. Radhakrishna 1971 Taxonomy in Anthropology. In Mathematics in Archaelogical and Historical Sciences, pp. 19–20. Edin. Univ. Press.

Rao, C. Radhakrishna 1977 Cluster Analysis Applied to a Study of Race Mixture in Human Populations. In Proceedings Michigan University Symposium, pp. 175–197.

Rao, C. Radhakrishna 1982a Diversity and Dissimilarity Coefficients: A Unified Approach. Theoret. Pop. Biol. 21: 24–43.

Rao, C. Radhakrishna 1982b Diversity: Its Measurement, Decomposition, Apportionment and Analysis. Sankhyā A, 44: 1–22.

Rao, C. Radhakrishna 1982c Gini-Simpson Index of Diversity: A Characterization, Generalization and Applications. Utilitas Mathematica 21B: 273–282.

Rao, C. Radhakrishna and R. Boudreau 1984 Diversity and Cluster Analyses of Blood Group Data on Some Human Populations. In The Human Population Genetics: The Pittsburgh Symposium, A. Chakravarti, ed., pp. 331–362. Van Nostrand Reinhold Co., New York.

Rao, C. Radhakrishna 1984 Use of Diversity and Distance Measures in the Analysis of Qualitative Data. In Multivariate Statistical Methods in Physical Anthropology, G. N. van Vark and W. W. Howells, eds., pp. 43–67. Reidel Publishing Company.

Torgerson, W. S. 1952 Multidimensional Scaling. Wiley, New York.

A New Statistical Test for the Comparison of the Standardized Differences Corresponding to Two Different Measurements: The Case of Petralona and Kabwe

G. N. VAN VARK and W. H. M. AMESZ-VOORHOEVE

Department of Anatomy
University of Groningen
Oostersingel 69
9713 EZ Groningen
The Netherlands

A. G. M. STEERNEMAN

Department of Econometrics
University of Groningen
P.O. Box 800
9700 AV Groningen
The Netherlands

and

D. W. READ

Department of Anthropology
University of California
405, Hilgard Avenue
Los Angeles, CA 90024
U.S.A.

ABSTRACT: A statistical test is presented that allows investigating the parts of e.g. the cranium for which discrimination between samples and individual specimens is most pronounced. The testing procedure uses pairwise comparison of variables to assess the relative contribution of each variable to this discrimination. The procedure is illustrated by comparing the Petralona and Kabwe crania with cranial samples of recent *Homo sapiens*, Neanderthals, and Asiatic *Homo erectus*.

KEY WORDS: crania, hominids, multivariate analysis

1. INTRODUCTION

The use of mathematical multivariate statistical methods in palaeoanthropology is still controversial. There are several reasons for this (Van Vark